

# Geometric Insights into Support Vector Machine Behavior using the KKT Conditions

**Iain Carmichael**

*Department of Statistics and Operations Research  
University of North Carolina  
Chapel Hill, NC 27516, USA*

IAIN@UNC.EDU

**J.S. Marron**

*Department of Statistics and Operations Research  
University of North Carolina  
Chapel Hill, NC 27516, USA*

MARRON@UNC.EDU

**Editor:**

## Abstract

The Support Vector Machine (SVM) is a powerful and widely used classification algorithm. Its performance is well known to be impacted by a tuning parameter which is frequently selected by cross-validation. This paper uses the Karush-Kuhn-Tucker conditions to provide rigorous mathematical proof for new insights into the behavior of SVM in the large and small tuning parameter regimes. These insights provide perhaps unexpected relationships between SVM and naive Bayes and maximal data piling directions. We explore how characteristics of the training data affect the behavior of SVM in many cases including: balanced vs. unbalanced classes, low vs. high dimension, separable vs. non-separable data. These results present a simple explanation of SVM's behavior as a function of the tuning parameter. We also elaborate on the geometry of complete data piling directions in high dimensional space. The results proved in this paper suggest important implications for tuning SVM with cross-validation.

**Keywords:** support vector machine, high-dimensional data, KKT conditions, data piling

## 1. Introduction

The *Support Vector Machine* (SVM) (see Schölkopf and Smola 2002; Shawe-Taylor and Cristianini 2004; Steinwart and Christmann 2008; Mohri et al. 2012; Murphy 2012 for a good overview) and its many variants is one of the most widely used and well studied classification algorithms. Classical classification algorithms, such as *logistic regression* and *Fisher linear discrimination* (FLD) are motivated by fitting a statistical distribution to the data. Hard margin SVM on the other hand is motivated by a geometric heuristic that leads directly to an optimization problem: maximize the margin between two classes of separable data. Soft margin SVM balances two competing objects; maximize the margin while penalizing points on the wrong side of the margin.

Like many optimization based classifiers SVM comes with a tuning parameter,  $C$ , that needs to be set. Different values of the tuning parameter can dramatically change SVM's behavior. The tuning parameter controls the trade off between the two terms in SVM's objective function (margin vs. slack). Unlike classifiers such as FLD, it can be challenging

to reason geometrically about how the position of the data points affects the classification normal vector and intercept. As discussed in standard texts such as the above books and Hsu et al. (2003) SVM tuning is frequently treated as black box optimization; the optimal value of  $C$  is typically selected via cross-validation.

The setting of this paper is the two class classification problem. We first consider linear classifiers then discuss important implications for kernel SVMs. We consider a wide range of data analytic regimes including: high vs. low dimension, balanced vs. unbalanced class sizes and separable vs. non-separable data.

A linear classifier is defined via the *normal vector* to its discriminating hyperplane and an *intercept* (or *offset*). The key idea in this paper is to compare *directions* of linear classifiers. Comparing the direction between two classifiers means comparing their normal vector directions; we say two directions are equivalent if one is a scalar multiple of the other (see Section 2). Note that two classifiers may have the same direction, but lead to different classification algorithms (i.e. the intercepts may differ).

Several of the classifiers discussed in this paper are not always defined, for example, hard margin SVM is only defined when the data are separable. We study these algorithms in restricted settings to gain insight into the more general setting.

It is common to mean center and scale each variable by its standard deviation before fitting a classifier (so called *standardization*). The *naive Bayes* classifier Friedman et al. (2001) is equivalent to first standardizing the data then computing the *mean difference* (MD). All theorems in this paper are proved for MD, however, they apply for naive Bayes if the data are first standardized. Therefore, all connections between SVM and MD imply connections between SVM and naive Bayes.

This paper provides new insight that explains commonly observed behavior of the SVM normal vector in the small  $C$  and in the large  $C$  regimes. We use the *Karush-Kuhn-Tucker* (KKT) conditions to characterize SVM's behavior in various settings and demonstrate connections between SVM and two other linear classifiers: the MD and the *maximal data piling direction* (MDP), Ahn and Marron (2010). In the small  $C$  regime soft margin SVM behaves like the MD classifier or a *cropped* version of the MD. In the large  $C$  regime SVM is related to the maximal MDP in high dimensional settings. The dimensionality of the space and the class sizes (balanced vs. unbalanced) have strong impacts on SVM's behavior. The main results are Theorem 6 and Corollary 8 for hard margin SVM and Theorems 17, 18 for soft margin SVM and Theorem 4 for data piling. In Section 5 these results are used to provide a number of new guidelines for tuning SVM with cross-validation.

## 1.1 Illustrative Example in Two Dimensions

Figures 1 and 2 illustrate some of the perhaps surprising SVM tuning behavior explained in this paper. The data in both figures are generated from a two dimensional Gaussian with identity covariance such that the class means are 4 apart. In Figure 1 the classes are balanced. The points in Figure 2 are the same points as the first figure, but one additional point is added to the positive class (blue squares) so the classes are unbalanced. In both cases the classes are linearly separable. These figures show the results of fitting SVM for a range of values of  $C$ .

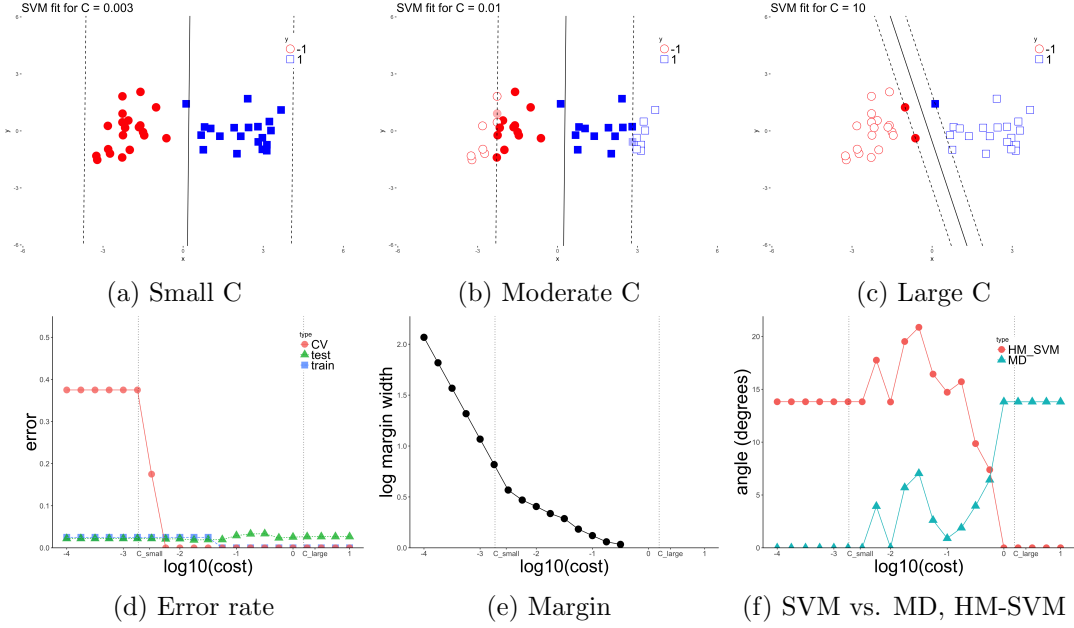


Figure 1: Balanced classes: The top panels show the SVM fit for a wide range of the tuning parameter  $C$  for a two dimensional data set. The bottom panels show performance measures as a function of  $C$ . Figure 1f uses angles between normal vectors to show SVM is equivalent to MD for small enough  $C$  and to hard margin SVM for large enough  $C$ . It also shows how cross-validation can be very different from the training and test error (1d). See the discussion in Section 5 for further details.

In both figures the data are shown in each panel in the top row. The SVM separating hyperplanes for a range of  $C$  values are also shown as solid lines with dashed lines indicating the SVM marginal hyperplanes. Filled in symbols are *support vectors* (see above SVM references).

As  $C$  shrinks to zero the SVM margin grows to infinity as suggested in the top left panels of both Figures (1a, 2a). This point is even more clear from the center bottom panels (1e, 2e) showing the margin width as a function of  $C$ . We will show in Theorem 17 that if the two classes are balanced then the SVM direction is equivalent to the mean difference direction for  $C < C_{\text{small}}$  (defined in Section 4). This phenomenon is demonstrated in the lower right panel of Figure 1f which shows the angle between the SVM and MD directions as a function of  $C$ . A formula for a threshold  $C_{\text{small}}$ , which is a function of the diameter of the training data, that gives this MD like behavior is given in Definition 15. The connection between soft margin SVM and MD in the small  $C$  regime for balanced data is proved in Hastie et al. (2004), however, they do not discuss the connection to the MD classifier, provide the value of  $C_{\text{small}}$  or give the more general results for the unbalanced case. This important bound,  $C_{\text{small}}$  is shown as a vertical dashed line in each bottom panel in Figures 1 and 2.

If the two classes are unbalanced then for  $C < C_{\text{small}}$  (the same function of the data as above) the SVM direction must satisfy constraints that make it a cropped mean difference direction; these constraints are given in Theorem 18 and Lemma 26. Figure 2f shows the

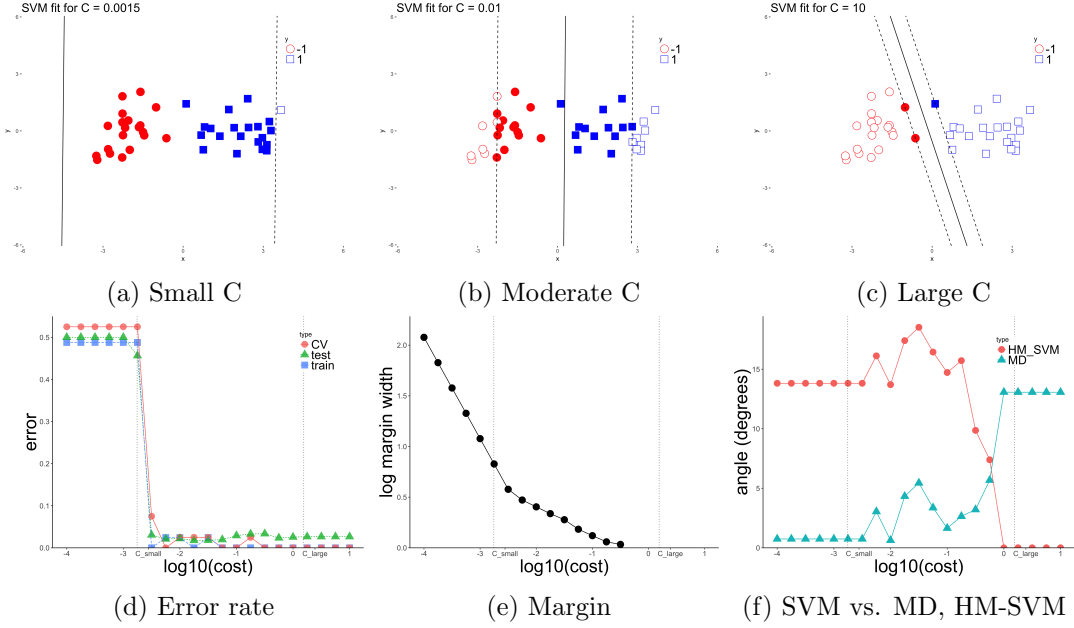


Figure 2: Unbalanced classes: the panels are the same as in Figure 1, but the data now have one additional point added. SVM comes very close to MD in the small  $C$  regime, but is not equivalent (2f). Unlike for the balanced case, cross-validation, train and test error all behave similarly (2d).

angle between the SVM and MD directions is small, but not zero for small  $C$ . Furthermore, for  $C < \frac{1}{2}C_{\text{small}}$  every training point is classified to the larger of the two classes as shown in Figures 2a and 2d. In particular the separating hyperplane (solid line) in Figure 2a is to the left of all data points (the left margin is pushed far away).

The results for the small  $C$  regime explain observed quirks of SVM. For example, the data in Figure 2 are the same as in Figure 1 but with one additional point added. The training and test error curves as a function of  $C$  shown in Figures 1d and 2d, however, look dramatically different. If the classes are unbalanced then as  $C$  continues to shrink to zero all training points are classified to the larger class. If the data are balanced then this behavior may not occur. These insights are important for improving the performance of SVM when tuned by cross-validation as discussed in Section 5.

In the large  $C$  regime soft margin SVM behaves like hard margin SVM or comes as close as it can. When the data are separable soft margin SVM becomes exactly hard margin SVM for large enough  $C > C_{\text{large}}$  as illustrated in the bottom right panels, Figures 1f and 2f. A specific formula for  $C_{\text{large}}$ , depending on the gap between classes in the training data, is given in Definition 16. This bound is indicated as a vertical dotted line in all lower panels. Hard margin SVM can have many points lying on the marginal hyperplanes leading to data piling, Marron et al. (2007). In some cases hard margin SVM leads to complete data piling; all training points in the same class are projected onto a single point by the normal vector. We give a geometric condition in Theorem 6 characterizing when hard margin SVM has

complete data piling. We also elaborate on the geometry of complete data piling directions. See Sections 5.3 and 5.4 for a detailed discussion of data piling.

The results in this paper provide a more complete understanding of hard margin SVM in high dimensions and of soft margin SVM for small and large values of the cost parameter. Soft margin SVM's behavior can change depending on characteristics of the data: balanced vs. unbalanced classes, whether  $d \geq n - 1$ , the two class diameter, whether the classes are separable and the gap between the two classes when they are separable. These insights have consequences for cross-validation procedures and will help the practitioner understand how SVM behaves and why it is succeeding/failing with real data sets.

This paper is organized as follows. Section 2 provides background information and notation. Section 3 discusses hard margin SVM in high dimensional settings. Section 4 discusses soft margin SVM in different cost parameter regimes. Section 5 elaborates on the consequences of the main results for data piling in high dimensions and for tuning SVM by cross-validation.

## 2. Setup and Notation

Suppose we have  $n$  labeled data points  $\{(\mathbf{x}_i, y_i)_{i=1}^n\}$  and index sets  $I_+, I_-$  such that  $y_i = 1$  if  $i \in I_+$ ,  $y_i = -1$  if  $i \in I_-$  and  $\mathbf{x}_i \in \mathbb{R}^d$ . Let  $n_+ = |I_+|$  and  $n_- = |I_-|$  be the class sizes (i.e.  $n_- + n_+ = n$ ).

We consider linear classifiers whose decision function is given by

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b,$$

where  $\mathbf{w} \in \mathbb{R}^d$  is the normal vector and  $b \in \mathbb{R}$  is the intercept (classification rule  $\text{sign}(f(x))$ ).

Given two vectors  $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$  we consider their *directions to be equivalent* if there exists  $a \in \mathbb{R}, a \neq 0$  such that  $a\mathbf{w} = \mathbf{v}$  (and we will write  $\mathbf{w} \propto \mathbf{v}$ ). Using this equivalence relation we can quotient  $\mathbb{R}^d$  into the space of directions (formally real projective space). Intuitively, this is the space of lines through the origin.

In this paper we consider the following linear classifiers: hard margin SVM, soft margin SVM (which we refer to as SVM), mean difference (also called *nearest centroid*), and the maximal data piling direction.

### 2.1 Linear Classifiers

The MD classifier selects the hyperplane that lies half way between the two class means. In particular the vector  $\mathbf{w}_{md}$  is given by the difference of the class means

$$\begin{aligned} \mathbf{w}_{md} &= \frac{1}{n_+} \sum_{i \in I_+} \mathbf{x}_i - \frac{1}{n_-} \sum_{i \in I_-} \mathbf{x}_i \\ &:= \bar{\mathbf{x}}_+ - \bar{\mathbf{x}}_-. \end{aligned} \tag{1}$$

Note that if the variables are first mean centered then scaled by the standard deviation then the mean difference is equivalent to the *naive Bayes* classifier.

For linear classifiers one frequently projects the data onto the one dimensional subspace spanned by the normal vector. *Data piling*, first discussed by Marron et al. (2007), is when multiple points have the same projection on the line spanned by the normal vector. For

example, all points on SVM's margin have the same image under the projection map. Ahn and Marron (2010) showed that when  $d \geq n - 1$  there are directions such that each class is projected to a single point i.e. there is *complete data piling*.

**Definition 1** A vector  $\mathbf{w} \in \mathbb{R}^d$  gives complete data piling for two classes of data if there exist  $a, b \in \mathbb{R}$ , with  $a \neq 0$  such that

$$\mathbf{w}^T \mathbf{x}_i = ay_i + b \text{ for each } i = 1, \dots, n,$$

where  $b$  is the midpoint of the projected classes and  $a$  is half the distance between the projected classes.

The MDP direction, as its name suggests, searches around all directions of complete data piling and finds the one that maximizes the distance between the two projected class images. This classifier has been studied in a number of papers such as Ahn et al. (2012), Lee et al. (2013), and Ahn and Marron (2010). The MDP direction can be computed analytically

$$\mathbf{w}_{mdp} = \hat{\Sigma}^-(\bar{\mathbf{x}}_+ - \bar{\mathbf{x}}_-),$$

where  $A^-$  is the Moore-Penrose inverse of a matrix  $A$  and  $\hat{\Sigma} = \frac{1}{n-1}(X - \bar{X})^T(X - \bar{X})$  is the global sample covariance matrix.

The MDP direction has an interesting relationship to Fisher linear discrimination. Recall the formula for FLD is

$$\mathbf{w}_{fld} = \hat{\Sigma}_{pool}^-(\bar{\mathbf{x}}_+ - \bar{\mathbf{x}}_-),$$

letting  $X_-$  and  $X_+$  be the data matrix for the respective classes and the *pooled sample covariance* is  $\hat{\Sigma}_{pool} = \frac{1}{n-2} [(X_+ - \bar{X}_+)^T(X_+ - \bar{X}_+) + (X_- - \bar{X}_-)^T(X_- - \bar{X}_-)]$  (in contrast with the global sample covariance in MDP). Ahn and Marron (2010) showed that in low dimensional settings FLD and the MDP formula are the same (though in low dimensional settings MDP does not give complete data piling); when  $d < n - 1$  the above two equations are equivalent, Ahn and Marron (2010).

Another view of this relation comes from the optimization perspective. FLD attempts to find the direction that maximizes the ratio of the projected “between-class variance to the within-class variance,” Bishop (2006). This problem is well defined only in low dimensions; in high dimensions when  $d \geq n - 1$  there exist directions of complete data piling where the within class projected variance is zero. In the high dimensional setting MDP searches around these directions of zero within class variance to find the one that maximizes the distance between the two classes (i.e. the between-class variance).

## 2.2 Support Vector Machine

Hard margin support vector machine is only defined when the data are linearly separable; it seeks to find the direction that maximizes the margin separating the two classes. It is defined as the solution to the following optimization problem,

$$\begin{aligned} & \underset{\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}}{\text{minimize}} && \frac{1}{2} \|\mathbf{w}\|^2 \\ & \text{subject to} && y_i(\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1, \text{ for } i = 1, \dots, n. \end{aligned} \tag{2}$$

When the data are not separable Problem (2) can be modified to give soft margin SVM by adding a tuning parameter  $C$  and slack variables  $\xi_i$  which allow points to be on the wrong side of the margin,

$$\begin{aligned} & \underset{\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}}{\text{minimize}} && \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \xi_i \\ & \text{subject to} && y_i(\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1 - \xi_i, \text{ for } i = 1, \dots, n \\ & && \xi_i \geq 0, \text{ for } i = 1, \dots, n. \end{aligned} \tag{3}$$

In both cases the direction is a linear combination of the training data points

$$\mathbf{w}_{svm} = \sum_{i \in I_+} \alpha_i \mathbf{x}_i - \sum_{i \in I_-} \alpha_i \mathbf{x}_i.$$

It turns out this linear combination always gives a direction that points between the convex hull of the two classes (see Definition 2).

### 3. Hard margin SVM in High Dimensions

We characterize the relationship between hard margin SVM, MD and MDP in high dimensions and provide novel insights into the geometry of complete data piling. In this section we assume  $d \geq n - 1$ . We further assume the data are in general position and separable, which implies the data are linearly independent if  $d \geq n$  and affine independent if  $d = n - 1$ . The data are in general position with probability 1 if they are generated by an absolutely continuous distribution in high dimensions. Typically the phenomena studied here happens in the  $n - 1$  dimensional affine space generated by the data.

Using geometric insights derived from the KKT conditions we give conditions for when

- the hard margin SVM direction is equivalent to the MDP direction,
- all three of the hard margin SVM, MDP and MD directions are equivalent.

Furthermore, we show that if the MD direction is equivalent to the MDP direction then they are both equivalent to the hard margin SVM direction. This section is organized as follows: we discuss data piling, we state the main results for data piling then for the mean difference, provide the KKT conditions, and then prove the main results.

#### 3.1 Hard Margin SVM and Complete Data Piling

We first define two important sets of directions.

**Definition 2** *Let  $B$  denote the set of all vectors associated with the directions that go between the convex hulls of the two classes i.e.*

$$B = \{a(\mathbf{c}_+ - \mathbf{c}_-) \mid a \in \mathbb{R}, a \neq 0, \text{ and } \mathbf{c}_j \in \text{conv}(\{\mathbf{x}_i\}_{i \in I_j}), j = \pm\}.$$

The set  $B$  may be all of  $\mathbb{R}^d$  if, for example, the two convex hulls intersect. When the data are linearly separable  $B$  is a strict subset of  $\mathbb{R}^d$ . Using ideas from Definition 1,

**Definition 3** Let  $P$  denote the vectors associated with directions that give complete data piling i.e.

$$P = \{\mathbf{v} \in \mathbb{R}^d | \exists a, b \in \mathbb{R}, a \neq 0 \text{ s.t. } \mathbf{v}^T \mathbf{x}_i = a \cdot y_i + b \text{ for each } i = 1, \dots, n\}.$$

Ahn and Marron (2010) point out there are infinitely many directions in the ( $n$  dimensional) subspace generated by the data that give complete data piling; in fact there is a great circle of directions in this subspace (if we parameterize directions by points on the unit sphere). Lemma 4 shows there is a single complete data piling direction that is also within the  $(n - 1)$  dimensional affine hull of the data. The remaining directions in  $P$  come from linear combinations of this unique direction in the affine hull and any vector normal to that hull.

**Theorem 4** The set of complete data piling directions,  $P$ , intersects the affine hull of the data in a single direction which is the maximal data piling direction.

Theorem 4 is proved in the appendix. A simple corollary of this theorem is:

**Corollary 5** The intersection of the between directions,  $B$ , and the complete data piling directions,  $P$ , is either empty or a single direction i.e.

$$B \cap P = \emptyset \text{ or } B \cap P = \{a\mathbf{v} | a \in \mathbb{R}\}.$$

The core results for hard margin SVM are summarized in the following theorem.

**Theorem 6** The hard margin SVM and MDP directions are equivalent if and only if there is a non-empty intersection between the between directions,  $B$ , and the complete data piling directions,  $P$ , i.e.

$$\mathbf{w}_{hm-svm} \propto \mathbf{w}_{mdp} \iff P \cap B \neq \emptyset$$

If this intersection is non-empty then it is a single direction and  $\mathbf{w}_{hm-svm} \in B \cap P$ .

This theorem also characterizes when SVM has complete data piling. Theorem 6 is a consequence of Corollary 5, Lemma 11, Lemma 12 and the KKT conditions.

An alternative way of deciding if  $B \cap P = \emptyset$  and computing the intersection if it exists is through the following linear program (proof of Theorem 7 is a straightforward exercise in linear programming).

**Theorem 7**  $B \cap P \neq \emptyset$  if and only if there is a solution to the following linear program

$$\begin{aligned} & \underset{\alpha \in \mathbb{R}^n_+, \beta \in \mathbb{R}^n_-, \mathbf{v} \in \mathbb{R}^d, b \in \mathbb{R}}{\text{minimize}} && 1 \\ & \text{subject to} && X\mathbf{v} + \mathbf{1}_n b = \mathbf{y} \\ & && \sum_{i \in I_+} \alpha_i \mathbf{x}_i - \sum_{i \in I_-} \beta_i \mathbf{x}_i = \mathbf{v} \\ & && \sum_{i \in I_+} \alpha_i = 1 \\ & && \sum_{i \in I_-} \beta_i = 1 \\ & && \alpha_i, \beta_i \geq 0 \text{ for } i = 1, \dots, n. \end{aligned} \tag{4}$$

In the case a solution  $\mathbf{v}$  exists then  $\mathbf{v} \in B \cap P$ .



The vector  $\mathbf{1}_n \in \mathbb{R}^n$  is the vector of ones,  $X$  is the  $\mathbb{R}^{n \times d}$  data matrix and  $\mathbf{y} \in \mathbb{R}^n$  is the vector of class labels. The first constraint says  $\mathbf{v}$  must be a complete data piling direction,  $\mathbf{v} \in P$ . The remaining constraints say  $\mathbf{v}$  must be a between direction,  $\mathbf{v} \in B$ .

### 3.2 Hard Margin SVM and the Mean Difference

The main result of this section characterizes when the MD, MDP and hard margin SVM directions all align. The proof of this result can be easily generalized from the mean difference to a wider class of directions. The key insight is that hard margin SVM gives complete data piling when it puts non-zero weight on all support vectors.

**Corollary 8** *The hard margin SVM and MD directions are equivalent if and only if all three of hard margin SVM, MD and MDP are equivalent i.e.*

$$\mathbf{w}_{hm-svm} \propto \mathbf{w}_{md} \iff \mathbf{w}_{md} \propto \mathbf{w}_{mdp}.$$

Corollary 8 is an immediate consequence of the following Theorem 10. Let  $T$  be the *strict between* directions: the subset of  $B$  where every point in both classes receives non-zero weight. For example, the MD puts positive weight on all training points so  $\mathbf{w}_{md} \in T$ .

**Definition 9** *Let  $T$  denote the vectors associated with the directions that go between the strict convex hulls of the two classes i.e.*

$$T = \{a(\mathbf{c}_+ - \mathbf{c}_-) \mid a \in \mathbb{R}, a \neq 0, \mathbf{c}_j = \sum_{i \in I_j} \lambda_i \mathbf{x}_i \text{ s.t. } \sum_{i \in I_j} \lambda_i = 1 \text{ and } \lambda_i > 0, i = 1, \dots, n, j = \pm\}.$$

**Theorem 10** *If the data are in general position then*

$$\mathbf{w}_{hm-svm} \in T \iff T \cap P \neq \emptyset.$$

*If this intersection is non-empty then it is a single direction and  $\mathbf{w}_{hm-svm} \propto \mathbf{w}_{mdp} \in T \cap P$ .*

### 3.3 Hard Margin KKT Conditions

Derivation and discussion of the KKT conditions can be found in Mohri et al. (2012). From the Lagrangian of Problem (2) we can derive the KKT conditions

$$\mathbf{w}_{hm-svm} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i, \tag{5}$$

$$\sum_{i=1}^n \alpha_i y_i = 0, \tag{6}$$

$$\alpha_i = 0 \text{ or } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1, \tag{7}$$

with  $\alpha_i \geq 0$  for each  $i = 1, \dots, n$ .

Condition (6) says that the sum of the weights in both classes has to be equal. Combining this with (5) we find that the hard margin SVM direction is given by

$$\mathbf{w}_{hm-svm} \propto \sum_{i \in I_+} \frac{\alpha_i}{A} \mathbf{x}_i - \sum_{i \in I_-} \frac{\alpha_i}{A} \mathbf{x}_i, \quad (8)$$

where  $\sum_{i \in I_+} \alpha_i = \sum_{i \in I_-} \alpha_i := A$ . Thus  $\mathbf{w}_{hm-svm} \in B$  i.e. the hard margin SVM direction is always a between direction. Equation (9.14) from Pham (2010) makes this clear and moreover shows that hard margin SVM is equivalent to finding the nearest points in the convex hulls of the two classes.

The last KKT condition (7) says that a point  $\mathbf{x}_i$  either lies on one of the marginal hyperplanes  $\{\mathbf{x} | \mathbf{w}_{hm-svm}^T \mathbf{x} = \pm 1\}$  or receives zero weight. In the former case when  $\alpha_i \neq 0$ ,  $\mathbf{x}_i$  is called a support vector.

The margin  $\rho$  is defined as the minimum distance from a training point to the separating hyperplane;  $\rho$  is also the orthogonal distance from the marginal hyperplanes to the separating hyperplane. The margin width is given by the magnitude of the normal vector

$$\rho^2 = \frac{1}{\|\mathbf{w}_{hm-svm}\|_2^2} = \frac{1}{\sum_{i=1}^n \alpha_i} := \frac{1}{\|\alpha\|_1}. \quad (9)$$

### 3.4 Proofs for Hard Margin SVM

The following lemma about SVM and MDP is a consequence of the KKT conditions.

**Lemma 11** *If hard margin SVM has complete data piling then the SVM direction is equivalent to the MDP direction i.e.*

$$\mathbf{w}_{hm-svm} \in P \implies \mathbf{w}_{hm-svm} \propto \mathbf{w}_{mdp}.$$

**Lemma 12** *If  $P \cap B \neq \emptyset$  then  $\mathbf{w}_{svm} \in P \cap B$ .*

**Proof** Let  $\mathbf{v} \in P \cap B$ . We show  $\mathbf{v}$  satisfies the KKT conditions. The lemma then follows since the KKT conditions necessary and sufficient for hard margin SVM (the constraints are qualified, see chapter 4 of Mohri et al. 2012).

Since  $\mathbf{v} \in B$  we have that  $\mathbf{v} \propto \mathbf{c}_+ - \mathbf{c}_-$  where  $\mathbf{c}_j \in \text{conv}(\{\mathbf{x}_i\}_{i \in I_j})$ . For some constant  $a > 0$

$$\mathbf{v} = a \left( \sum_{i \in I_+} \lambda_i \mathbf{x}_i - \sum_{i \in I_-} \lambda_i \mathbf{x}_i \right),$$

where

$$\sum_{i \in I_+} \lambda_i = \sum_{i \in I_-} \lambda_i = 1 \text{ and } \lambda_i \geq 0.$$

Since  $\mathbf{v} \in P$  we can select  $b, \mathbf{v}$  such that

$$y_i(\mathbf{x}_i \cdot \mathbf{v} + b) = 1 \quad \forall i.$$

But these three equations are the KKT conditions with  $\alpha_i = a\lambda_i$ .

■

### Proof of Theorem 10

If  $T \cap P \neq \emptyset$  then Theorem 6 applies. If  $\mathbf{w}_{hm-svm} \in T$  we show that every point must be a support vector. General position is a key assumption.

Since  $\mathbf{w}_{hm-svm} \in T$  there exists some  $a > 0$  such that

$$\mathbf{w}_{hm-svm} = a \left( \sum_{i \in I_+} \pi_i \mathbf{x}_i - \sum_{i \in I_-} \eta_i \mathbf{x}_i \right),$$

where  $\sum_{i \in I_+} \pi_i = \sum_{i \in I_+} \eta_i = 1$  and  $\pi_i, \eta_i > 0$  for each  $i$ . Using the KKT conditions  $\mathbf{w}_{hm-svm} = \sum_{i \in I_+} \alpha_i \mathbf{x}_i - \sum_{i \in I_-} \alpha_i \mathbf{x}_i$  and subtracting we get

$$\sum_{i \in I_+} (a\pi_i - \alpha_i) \mathbf{x}_i - \sum_{i \in I_-} (a\eta_i - \alpha_i) \mathbf{x}_i = 0.$$

Note that this is an affine combination since

$$\begin{aligned} \sum_{i \in I_+} (a\pi_i - \alpha_i) - \sum_{i \in I_-} (a\eta_i - \alpha_i) &= \left( \sum_{i \in I_+} \alpha_i - \sum_{i \in I_-} \alpha_i \right) + a \left( \sum_{i \in I_+} \pi_i - \sum_{i \in I_-} \eta_i \right) \\ &= A - A + 1 - 1 = 0. \end{aligned}$$

Since the data are affine independent we get that each of the above coefficients must be zero so

$$\alpha_i = \begin{cases} a\pi_i & \text{if } i \in I_+ \\ a\eta_i & \text{if } i \in I_- \end{cases}$$

In particular each coefficient  $\alpha_i \neq 0$  therefore each point is support vector and lies on the margin.

■

## 4. Soft Margin SVM Small and Large $C$ Regimes

This section characterizes the behavior of SVM for the small and large regimes of the cost parameter  $C$ . We make no assumptions about the dimension of the data  $d$ . We state the main results for the small and large  $C$  regimes, provide the KKT conditions, then prove the tuning regimes results.

We first make two geometric definitions that play an important role in characterizing SVM's tuning behavior. The two class *diameter* measures the spread of the data.

**Definition 13** *Let the two class diameter be*

$$D := \max_{\mathbf{x}_+ \in I_+, \mathbf{x}_- \in I_-} \|\mathbf{x}_+ - \mathbf{x}_-\|.$$

The *gap* measures the separation between the two data classes.

**Definition 14** *Let the two class gap  $G$  be the minimum distance between points in the convex hulls of the two classes i.e.*

$$G := \min_{\mathbf{c}_j \in \text{conv}(\{\mathbf{x}_i\}_{i \in I_j})} \|\mathbf{c}_+ - \mathbf{c}_-\|.$$

Using the above geometric quantities we define two threshold values of  $C$  which determine when the SVM enters its different behavior regimes.

**Definition 15** *For two classes of data let*

$$C_{\text{small}} := \frac{1}{2 \max(n_+, n_-) D^2}, \quad (10)$$

where  $D$  is the diameter of the training data.

**Definition 16** *If the two data classes are linearly separable let*

$$C_{\text{large}} := \frac{2}{G^2}, \quad (11)$$

where  $G$  is the gap between the classes.

As illustrated in Figures 1 and 2, the main result for the small  $C$  regime is

**Theorem 17** *When  $C < C_{\text{small}}$  given above in Definition 15*

- *If the classes are balanced then the SVM direction becomes the mean difference direction i.e.  $\mathbf{w}_{\text{svm}} \propto \mathbf{w}_{\text{md}}$ .*
- *If the classes are unbalanced then the SVM direction satisfies the constraints in Lemma 26 making it a cropped mean difference. As  $C$  continues to shrink and if the classes are unbalanced then the separating hyperplane goes to infinity and every data point is classified to the larger of the two classes.*

If the data are separable then in the large  $C$  regime soft margin SVM becomes equivalent to hard margin SVM for sufficiently large  $C$ .

**Theorem 18** *If the training data are separable then when  $C > C_{\text{large}}$  soft margin SVM is equivalent to the hard margin SVM solution i.e.  $\mathbf{w}_{\text{svm}} = \mathbf{w}_{\text{hm-svm}}$ .*

Note that  $C_{\text{small}}$  and  $C_{\text{large}}$  are lower and upper bounds—their respective limiting behavior may (and typically does) happen for  $C$  larger than  $C_{\text{small}}$  and  $C$  smaller than  $C_{\text{large}}$ .

#### 4.1 Soft Margin SVM KKT Conditions

The KKT conditions for soft margin SVM are (see Mohri et al. 2012 for derivations)

$$\mathbf{w}_{svm} = \sum_{i \in I_+} \alpha_i \mathbf{x}_i - \sum_{i \in I_-} \alpha_i \mathbf{x}_i, \quad (12)$$

$$\sum_{i \in I_+} \alpha_i = \sum_{i \in I_-} \alpha_i := A, \quad (13)$$

$$\alpha_i + \mu_i = C \text{ for } i = 1, \dots, n, \quad (14)$$

$$\alpha_i = 0 \text{ or } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1 - \xi_i \text{ for } i = 1, \dots, n, \quad (15)$$

$$\xi_i = 0 \text{ or } \mu_i = 0 \text{ for each } i, \quad (16)$$

with  $\alpha_i, \mu_i \geq 0$ .

For soft margin SVM we define the marginal hyper planes to be  $\{\mathbf{x} | \mathbf{x}^T \mathbf{w}_{svm} = \pm 1\}$  and the margin width (or just margin),  $\rho$  the distance from the separating hyperplane to the marginal hyperplanes. By construction  $\rho = \frac{1}{\|\mathbf{w}_{svm}\|}$ . For soft margin SVM the margin does not have the same meaning as in the hard margin case, but still plays an important role.

As with hard margin SVM the soft margin direction is always a between direction. Again points  $\mathbf{x}_i$  such that  $\alpha_i \neq 0$  are called support vectors. We further separate support vectors into two types.

**Definition 19** *Margin vectors are support vectors  $\mathbf{x}_i$  such  $\alpha_i \neq 0$  and  $\xi_i = 0$ .*

**Definition 20** *Slack vectors are support vectors  $\mathbf{x}_i$  such  $\alpha_i \neq 0$  and  $\xi_i > 0$ .*

Margin vectors are support vectors lying on one of the two marginal hyperplanes. Slack vectors are support vectors lying strictly on the inside of the marginal hyperplanes. Call the set of margin vectors in each class  $M_j$  and the set of slack vectors  $L_j$  for  $j = \pm$ .

The KKT conditions imply

- all support vectors receive weight upper bounded by  $C$  ( $\mathbf{x}_i \in M_j \implies 0 < \alpha_i \leq C$ )
- slack vectors receive weight exactly  $C$  ( $\mathbf{x}_i \in L_j \implies \alpha_i = C$ )

The following weight balance constraint holds

$$C|L_+| + \sum_{i \in M_+} \alpha_i = C|L_-| + \sum_{i \in M_-} \alpha_i. \quad (17)$$

Without loss of generality we assume that the positive class is the larger of the two classes i.e.  $n_+ \geq n_-$ . Unbalanced classes means  $n_+ > n_-$ .

## 4.2 Proofs for Small $C$ Regime

As  $C \rightarrow 0$  the margin width increases to infinity ( $\rho \rightarrow \infty$ ). As the margin width grows as many points as possible become slack vectors and all slack vectors get the same weight  $\alpha_i = C$ . Hence if the classes are balanced the SVM direction will be equivalent to the mean difference. If the classes are unbalanced then there will be some margin vectors which receive weight  $\alpha_i \leq C$ . The number of margin vectors is bounded by the class sizes and the dimension.

Note the diameter,  $D$ , does not change if we consider the convex hull of the two classes (proof of Lemma 21 is a straight forward exercise).

**Lemma 21**

$$\max_{\mathbf{c}_j \in \text{conv}(\{\mathbf{x}_i\}_{i \in I_j})} \|\mathbf{c}_+ - \mathbf{c}_-\| = \max_{\mathbf{x}_j \in I_+} \|\mathbf{x}_+ - \mathbf{x}_-\| =: D.$$

As  $C \rightarrow 0$  the magnitude of  $\mathbf{w}_{svm}$  goes to zero. In particular, the KKT conditions give the following bound.

**Lemma 22** *For a given  $C$  the magnitude of the SVM solution is*

$$\|\mathbf{w}_{svm}\| \leq n_+ C \cdot D.$$

**Proof** From the KKT conditions we have

$$\mathbf{w}_{svm} = \sum_{i \in I_+} \alpha_i \mathbf{x}_i - \sum_{i \in I_-} \alpha_i \mathbf{x}_i$$

and

$$\sum_{i \in I_+} \alpha_i = \sum_{i \in I_-} \alpha_i =: A.$$

Computing the magnitude of  $\mathbf{w}_{svm}$

$$\|\mathbf{w}_{svm}\| = A \left\| \sum_{i \in I_+} \frac{\alpha_i}{A} \mathbf{x}_i - \sum_{i \in I_-} \frac{\alpha_i}{A} \mathbf{x}_i \right\|.$$

Since the two terms are convex combinations we get

$$\mathbf{w}_{svm} \leq A \sup_{\mathbf{c}_j \in \text{conv}(\{\mathbf{x}_i\}_{i \in I_j})} \|\mathbf{c}_+ - \mathbf{c}_-\|.$$

applying Lemma 21

$$\begin{aligned} \mathbf{w}_{svm} &= A \max_{\mathbf{x}_j \in I_+} \|\mathbf{x}_+ - \mathbf{x}_-\| \\ \mathbf{w}_{svm} &= AD. \end{aligned}$$

Since  $0 \leq \alpha_i \leq C$  we get  $A \leq n_1 C$  thus proving the bound. ■

Since the magnitude of  $\mathbf{w}_{svm}$  determines the margin width, using the previous lemma we get the following corollary.

**Corollary 23** *The margin  $\rho$  goes to infinity as  $C$  goes to zero. In particular*

$$\rho = \frac{1}{\|\mathbf{w}_{svm}\|} \geq \frac{1}{n_+CD}.$$

Since the margin width increases, for small enough  $C$  the smaller class becomes all slack variables.

**Lemma 24** *If  $C < C_{small} := \frac{1}{2n_+D^2}$  then all points in the smaller class become slack vectors i.e.  $\xi_i > 0$  for all  $i \in I_-$ .*

**Proof** By Corollary 23 the margin width goes to infinity as  $C \rightarrow 0$  since

$$\rho \geq \frac{1}{n_+CD}.$$

Recall the margin width,  $\rho$ , is the distance from the separating hyperplane to the marginal hyperplanes. Note that if  $\rho > \frac{1}{2}D$  then at least one class must be complete slack. Thus if  $C < \frac{1}{2n_+D^2}$  at least one class must be complete slack i.e.  $\xi_i > 0$  for all  $i \in I_j$  for  $j = +$  and/or  $j = -$ . If the classes are balanced then either class can become complete slack (or both classes).

If the classes are unbalanced i.e.  $n_- < n_+$  then the smaller class becomes complete slack. To see this, assume for the sake of contradiction that the larger class becomes complete slack i.e.  $\xi_i \neq 0$  for each  $i \in I_+$ . Then the KKT conditions imply  $\alpha_i = C$  for each  $i \in I_+$ . KKT condition 13 says

$$\begin{aligned} \sum_{i \in I_+} \alpha_i &= \sum_{i \in I_-} \alpha_i \\ n_+C &= \sum_{i \in I_-} \alpha_i. \end{aligned}$$

But  $\alpha_i \leq C$  and  $n_- < n_+$  by assumption therefore this constraint cannot be satisfied.  $\blacksquare$

If the classes are balanced then the margin swallows both classes and the SVM direction becomes the mean difference direction.

**Lemma 25** *If the classes are balanced and  $C < C_{small}$  the SVM direction is equivalent to the mean difference direction i.e.  $\mathbf{w}_{svm} \propto \mathbf{w}_{md}$ .*

**Proof** When  $C < C_{small}$  one of the classes (without loss of generality the negative class) becomes slack i.e.  $\xi_i > 0$  for each  $i \in I_-$  thus  $\alpha_i = C$  for each  $i \in I_-$ . The KKT conditions then require

$$\sum_{i \in I_+} \alpha_i = \sum_{i \in I_-} \alpha_i = n_-C.$$

Since  $\alpha_i \leq C$  and  $|I_+| = n_-$  this constraint can only be satisfied if  $\alpha_i = C$  for each  $i \in I_+$ . We now have

$$\mathbf{w}_{svm} = \sum_{i \in I_+} C\mathbf{x}_i - \sum_{i \in I_-} C\mathbf{x}_i$$

$$\mathbf{w}_{svm} = C \frac{n}{2} (\bar{\mathbf{x}}_+ - \bar{\mathbf{x}}_-) \propto \mathbf{w}_{md}.$$

■

**Lemma 26** *If the classes are unbalanced and  $C < C_{small}$  the SVM solution satisfies the following constraint*

$$\mathbf{w}_{svm} = \sum_{i \in M_+} \alpha_i \mathbf{x}_i + C \sum_{i \in L_+} \mathbf{x}_i - C \sum_{i \in L_-} \mathbf{x}_i, \quad (18)$$

subject to

$$\sum_{i \in M_+} \alpha_i = C(|L_+| - n_-). \quad (19)$$

**Proof** Recall for  $C < C_{small}$  we have  $\xi_i > 0$  for  $i \in L_-$ . From the KKT conditions  $\xi_i > 0 \implies \mu_i = 0 \implies \alpha_i = 0$  meaning  $\alpha_i = C$  for each  $i \in L_-$ . The weight balance constraint 17 from the KKT conditions becomes

$$C|L_+| + \sum_{i \in M_+} \alpha_i = C|L_-| + \sum_{i \in M_-} \alpha_i,$$

which then implies the conditions on  $\mathbf{w}_{svm}$ . ■

Lemma 26 characterizes a kind of cropped mean difference. The mean difference direction points between the mean of the first class and the mean of the second class. Recall  $\mathbf{w}_{svm}$  always goes between points in the convex hulls of the two classes. Equation 18 says that in the small  $C$  regime  $\mathbf{w}_{svm}$  points between the mean of the negative (smaller) class (the third term) and a point that is close to the mean in the positive (larger) class. The cropping happens by ignoring non-support vectors. Additionally, points on the margin do not necessarily receive equal weight. However, Equation 19 bounds the amount of weight put on points on margin points.

**Corollary 27** *When  $C < C_{small}$  the positive (larger) class can have at most  $n_-$  slack vectors. If the larger class has more than  $n_-$  support vectors then at least one of them must be a margin vector.*

As  $C$  continues to shrink past  $C_{small}$  the margin width continues to grow. Eventually the separating hyperplane will be pushed past the smaller class and every training point will be classified to the larger class.

**Corollary 28** *If the classes are unbalanced and  $C < \frac{1}{2}C_{small}$  then every training point is classified to the positive (larger) class.*

If the data are in general position we can strengthen some of the above results.

**Lemma 29** *If the data are in general position the larger class can have at most  $n_- + d - 1$  support vectors.*



### 4.3 Proofs for Large $C$ Regime

**Lemma 30** *If there is at least one slack vector then for a given  $C$*

$$\|\mathbf{w}_{svm}\| \geq CG,$$

*or equivalently*

$$\rho \leq \frac{1}{CG},$$

*where  $G$  is the class gap.*

**Proof** From the KKT conditions

$$\|\mathbf{w}_{svm}\| = \left\| \sum_{i \in I_+} \alpha_i \mathbf{x}_i - \sum_{i \in I_-} \alpha_i \mathbf{x}_i \right\|,$$

$$\|\mathbf{w}_{svm}\| = A \left\| \sum_{i \in I_+} \frac{\alpha_i}{A} \mathbf{x}_i - \sum_{i \in I_-} \frac{\alpha_i}{A} \mathbf{x}_i \right\|,$$

where  $A = \sum_{i \in I_+} \alpha_i = \sum_{i \in I_-} \alpha_i$ . Since the two sums are convex combinations, using the definition of  $G$  we get

$$\|\mathbf{w}_{svm}\| \geq AG.$$

Since there is at least one slack vector there is at least one  $i$  such that  $\alpha_i = C$  thus  $A \geq C$  and the result follows.  $\blacksquare$

## 5. Discussion

In this section we first discuss consequences for SVM tuning, then the geometry of complete data piling, the maximal data piling direction, and then kernel SVM.

### 5.1 SVM Regimes

This paper provides a detailed understanding of SVM's behavior as a function of the tuning parameter and how its behavior is affected by various attributes of the data. For sufficiently large  $C > C_{\text{large}}$ , Theorem 18 shows that soft margin SVM is equivalent to hard margin SVM if the data are separable. For sufficiently small  $C < C_{\text{small}}$  Theorem 17 shows how the soft margin SVM is related to the mean difference.

Often the variables are standardized before fitting SVM (first mean centered then scaled by the standard deviation). In this case the mean difference classifier is equivalent to the naive Bayes classifier. Therefore, if the data are standardized all results about SVM and the mean difference hold for SVM and naive Bayes.

In the small  $C$  regime, if the data are balanced then the SVM direction becomes exactly the mean difference direction. Lemma 1 from Hastie et al. (2004) proves this result, but does not discuss the connection between SVM and the mean difference classifier. If the data are unbalanced then the SVM direction becomes a cropped mean difference direction as characterized in Lemma 26. Points far away from the separating hyperplane are ignored.

Points strictly within the margin receive the same weight. Points on the margin may receive smaller weight than the slack vectors. The number of margin vectors is bounded by the dimension. These relationships are made precise by Lemma 26 and Lemma 29.

Corollary 23 shows that as  $C$  shrinks the margin width blows up. When the margin blows up there are two cases. First case, both marginal hyperplanes go to infinity and the separating hyperplane does not. Second case, the separating hyperplane goes to infinity resulting in all data being assigned to a single (larger) class i.e. the hyperplane bounces off the larger class. Corollary 27 shows that the latter case always happens when the data are unbalanced. When the data are balanced either case can occur.

These insights explain the stark differences between the tuning error curves in Figures 1d and 2d. The cross-validation test error curve behaves dramatically different in Figure 1d than the training and test error curves in the small  $C$  regime. In this example the training classes are balanced while the cross-validation training folds are unbalanced. Therefore, as discussed above, the separating hyperplane bounces off the larger cross-validation training class and misclassifies all points in the smaller class.

## 5.2 Tuning SVM via Cross-Validation

We have shown that SVM's behavior in the small and large  $C$  regimes can depend on characteristics of the data:

- balanced vs. unbalanced classes,
- the two class diameter  $D$ ,
- whether or not  $d \geq n - 1$ ,
- whether or not the classes are separable,
- the gap between the two classes  $G$ .

Each of these characteristics can change between the full training set and the cross-validation training sets. When the characteristics change, so can SVM's behavior for small and large values of  $C$ . Therefore SVM may behave very differently for the cross-validation folds than for the full training data. One dramatic example of this change in behavior can be seen in Figure 1d as discussed in Section 5.1.

Another example of tuning behavior differences between the training and cross-validation data can be seen by looking carefully at Figure 2d. In this figure we can see the cross-validation error rate shoots up for larger values of  $C$  than the train/test error rates. The error increases dramatically for small values of  $C$  because of the margin explosion phenomena discussed in Section 5.1. The value of  $C_{\text{small}}$  that guarantees this behavior is a function of the two class diameter  $D$  (see Definition 15). Since there are fewer points in the cross-validation training set, the diameter is smaller meaning the value of  $C_{\text{small}}$  is larger causing the margin to explode for larger values of  $C$ .

Different data domains in terms of  $n \ll d$ ,  $n \sim d$ , and  $n \gg d$  can make the above characteristics more or less sensitive to change induced by subsampling. For example, if  $n \gg d$  then subsampling is least likely to change whether  $d \geq n - 1$  or significantly modify

the diameter  $D$ . With a kernel, however, even if the original  $n \gg d$  then it may no longer be true that  $n \gg d_{\text{implicit}}$  where  $d_{\text{implicit}}$  is the dimension of the implicit kernel space.

When  $n$  is larger than  $d$ , but not by much, then subsampling is likely to change whether or not  $d \geq n - 1$  and whether or not the data are separable. In this case the full training data may not be separable, but the cross-validation sets may be. This means large values of  $C$  will cause soft margin SVM to become hard margin SVM for cross-validation, but never for the full training data. This could result in the SVM direction being very different between cross-validation and training.

When  $d \geq n - 1$  soft margin SVM will become hard margin SVM for  $C \geq C_{\text{large}}$  which depends on the gap  $G$  between the two classes. Subsampling the data will cause this gap to increase meaning  $C_{\text{large}}$  decreases. In this case the hard margin behavior will occur for smaller values of  $C$  in the cross-validation sets than for the full training set.

It may be desirable to perform cross-validation in a way that is least likely to change some of the above characteristics between the full and the cross-validation training data set. For example, the following guidelines may be worth exploring.

- If the full training data are balanced one should ensure the cross-validation training classes are also balanced.
- Cross-validation with a large number of folds (e.g. leave one out CV) is presumably least likely to modify the above characteristics of the data.
- When  $n > d$  it could be judicious to make sure that  $n_{cv} > d$  for each cross-validation training set.

When the margin explosion behavior occurs it is really a problem with the intercept, not the SVM normal vector direction. Therefore, it may also be worth modifying the intercept term for small values of  $C$ . For example, for  $C < C_{\text{small}}$  one could fix the intercept such that the separating hyperplane lies halfway between the two class means.

When performing a grid search to find the optimal value of  $C$  one should restrict the search to  $C \in (C_{\text{small}}, C_{\text{large}})$  (note  $C_{\text{large}}$  may be infinity if the data are not separable). These bounds can be computed easily from a given data set.  $C_{\text{small}}$  based on the between class diameter  $D$ . If the data are separable,  $C_{\text{large}}$  is driven by the gap,  $G$ , between the convex hulls of the two classes and there are several efficient algorithms that solve this problem, Kaown (2009). Hastie et al. (2004) derives an efficient algorithm to find the SVM solution for every value of the cost parameter. When using this algorithm knowledge of  $C_{\text{small}}$  and  $C_{\text{large}}$  does not provide further computational benefits beyond this algorithm. However this restriction would prove useful if the user does not implement this entire cost path algorithm.

### 5.3 Geometry of Complete Data Piling

In this section we consider directions to be points on the unit sphere; the equivalence class of a single direction is represented by two antipodal points. When  $d \geq n$  there are an infinite number of directions  $P$  that give complete data piling. If we restrict ourselves to the  $n$  dimensional subspace generated by the data there are still an infinite number of directions that give complete data piling Ahn and Marron (2010); within this subspace  $P$  forms a great circle of directions.

Theorem 4 says that if we further restrict ourselves to the  $n - 1$  dimensional affine hull of the data there is only a single direction of complete data piling and this direction is the maximal data piling direction. The aforementioned great circle of directions intersects the subspace parallel to the affine hull of the data at two points (i.e. a single direction).

## 5.4 Data Piling

Hard margin SVM always has some data piling; support vectors in the same class project to the same point. If the data are in general position the number of support vectors is bounded by the dimension (Lemma 29). When  $d \geq n - 1$  Theorem 6 gives geometric conditions for when hard margin SVM gives complete data piling.

Typically, even when  $d \geq n - 1$ , hard margin SVM only has partial data piling. Complete data piling is a strict constraint and the SVM normal vector can usually wiggle away from the MDP direction to find a larger margin. Hard margin SVM can be viewed as a cropped MDP direction; the normal vector  $\mathbf{w}_{hm-svm}$  is the MDP direction of the subset of the data on the margin, points away from the margin are ignored.

This raises the question: is complete data piling with hard margin SVM a probability zero event when the data are generated by an absolutely continuous distribution? We suspect the answer is no: it occurs with positive, but typically small probability. For example consider three points in  $\mathbb{R}^2$ .

**Conjecture 31** *If the data are generated by an absolutely continuous distribution and  $d \geq n - 1$  then hard margin SVM has complete data piling with positive probability.*

Often data piling may not be desirable e.g. the normal vector may be sensitive to small scale noise artifacts Marron et al. (2007). Additionally, the projected data have a degenerate distribution since multiple data points lie on top of each other. However there are cases, such as an autocorrelated noise distribution, when the maximal data piling direction performs well, Miao (2015).

## 5.5 Kernels

Kernel SVM corresponds to linear SVM in a transformed data space so all of the results of this paper apply to kernel SVM. For example, in the small  $C$  regime SVM behaves like the kernel mean difference. One can explicitly compute  $C_{\text{small}}$  and  $C_{\text{large}}$  for difference kernels since these bounds depend only on the dot products between the data points.

Kernels map the data into a high dimensional space. For data where  $d < n - 1$  the corresponding kernel transformed data can be separable in the implicit space and/or that implicit space may have dimension larger than  $n - 1$ . Therefore the large  $C$  and hard margin results may be particularly relevant for kernel SVM.

## Acknowledgments

This research was supported in part by the National Science Foundation under Grant No. 1633074.

## Appendix A.

In this section we prove Theorem 4. Online supplementary material including code to reproduce the figures in this paper, proofs that were omitted for brevity and simulations can be found at: [https://github.com/idc9/svm\\_geometry](https://github.com/idc9/svm_geometry).

### Proof of Theorem 4

We first prove the existence and uniqueness of complete data piling directions  $P$  in the affine hull of the data. We then show that this unique, affine data piling direction is in fact the direction of maximal data piling.

Recall we assume that  $d \geq n - 1$  and the data are in general position. Let the set of affine directions  $A$  be given as follows

$$A = \{\mathbf{a}_1 - \mathbf{a}_2 | \mathbf{a}_j \in \text{aff}(\{\mathbf{x}_i\}_1^n), j = 1, 2\}.$$

Note that  $A$  is the  $n-1$  dimensional subspace parallel to the affine space  $\text{aff}(\{\mathbf{x}_i\}_1^n)$  generated by the data i.e.  $A$  contains the origin.

We first show that without loss of generality  $d = n - 1$ . Note that both  $A$  and  $P$  are invariant to a fixed translation of the data. Therefore, we may translate the data so that  $0 \in \text{aff}(\{\mathbf{x}_i\}_1^n)$  (e.g. translate by the mean of the data). The data now span an  $n - 1$  dimensional subspace since the affine hull of the data now contains the origin. Furthermore,  $\text{span}(\{\mathbf{x}_i\}_1^n) = \text{aff}(\{\mathbf{x}_i\}_1^n) = A$ . Thus without loss of generality we may consider the data to in fact be  $n - 1$  dimensional (i.e.  $d = n - 1$ ).

We are now looking for a vector  $\mathbf{v} \in A$  that gives complete data piling. Note by the above discussion and assumption we have  $A = \mathbb{R}^d$ . This means we are looking for  $\mathbf{v} \in \mathbb{R}^d$  and  $a, b \in \mathbb{R}$  with  $a \neq 0$  satisfying the following  $n$  linear equations

$$\mathbf{x}_i^T \mathbf{v} = ay_i + b \text{ for } i = 1, \dots, n.$$

Since the magnitude of  $\mathbf{v}$  is arbitrary we fix  $a = 1$  without loss of generality. We now have

$$\mathbf{x}_i^T \mathbf{v} = y_i + b \text{ for } i = 1, \dots, nm$$

which can be written in matrix form as

$$X\mathbf{v} + b\mathbf{1}_n = \mathbf{y} \tag{20}$$

where  $X \in \mathbb{R}^{n \times d}$  is the data matrix whose rows are the data vectors  $\mathbf{x}_i$  and  $\mathbf{y} \in \mathbb{R}^n$  is the vector of class labels. This is a system of  $n$  equations in  $\mathbb{R}^{d+1}$  which can be seen by appending 1 onto the end of each  $\mathbf{x}_i$  i.e.  $\tilde{\mathbf{x}}_i = (\mathbf{x}_i, 1) \in \mathbb{R}^{d+1}$  and letting  $\mathbf{w} = (\mathbf{v}, b)$ . Then Equation 20 becomes

$$\tilde{X}\mathbf{w} = \mathbf{y} \tag{21}$$

where  $\tilde{X} \in \mathbb{R}^{n \times d+1}$  is the appended data matrix.

Recall that we assumed  $d = n - 1$  so Equation 21 is a system of  $n$  equations in  $\mathbb{R}^n$ . Further recall that the data are in general position meaning that the  $n$  data points are affine independent in the  $n - 1$  dimensional subspace of the data. Affine independence is equivalent to linear independence of  $\{(\mathbf{x}_i, 1)\}_1^n$ . Therefore the matrix  $\tilde{X} \in \mathbb{R}^{n \times n}$  has full rank and Equation 21 always has a solution,  $\mathbf{v}^*$ , and this solution is unique.

Existence of a solution to Equation 21 shows that  $P \cap A \neq \emptyset$ . Uniqueness of the solution to Equation 21 shows that this intersection  $P \cap A$  can have only one direction of which  $\mathbf{v}^*$  is a representative element.

We now show that  $\mathbf{v}^*$  is in fact the maximal data piling direction. We no longer assume that  $d = n - 1$ .

We first construct an orthonormal basis  $\{\mathbf{t}_i\}_1^d$  of  $\mathbb{R}^d$  as follows. Let the first  $n - 1$  basis vectors  $\mathbf{t}_1, \dots, \mathbf{t}_{n-1}$  span  $A$ . Let  $\mathbf{t}_n$  be orthogonal to  $A$  but in the span of the data  $\{\mathbf{x}_i\}_1^n$  (recall the data span an  $n$  dimensional space while the affine hull of the data is  $n - 1$  dimensional). Let the remaining  $d - n + 1$  basis vectors be orthogonal to  $A$  and the span of the data.

We show that the vector  $\mathbf{t}_n$  projects every data point onto a single point i.e.  $\mathbf{x}_i^T \mathbf{t}_n = c$  for each  $i = 1, \dots, n$  and some  $c \in \mathbb{R}$ . Suppose we translate  $\text{aff}(\{\mathbf{x}_i\}_1^n)$  along  $\mathbf{t}_n$  until the origin lies in the affine hull of the translated data. In particular, the data now span an  $n - 1$  dimensional subspace that is orthogonal to  $\mathbf{t}_n$  (where as before they spanned an  $n$  dimensional subspace). We now have that for some  $c \in \mathbb{R}$

$$\mathbf{t}_n^T (\mathbf{x}_i + c\mathbf{t}_n) = 0 \text{ for each } i = 1, \dots, n$$

$$\mathbf{t}_n^T \mathbf{x}_i = c \text{ for each } i = 1, \dots, n$$

since  $\mathbf{t}_n$  is unit norm.

Let  $\mathbf{v} \in \mathbb{R}^d$  be a representative vector of the direction in the affine hull of the data that gives complete data piling (given above). Suppose  $\mathbf{v}$  has unit norm and is oriented such that

$$\mathbf{v}^T \mathbf{x}_i = ay_i + b$$

for some  $a, b \in \mathbb{R}$  with  $a > 0$  (note fixing  $a > 0$  eliminates the antipodal symmetry of data piling vectors).

We now show that  $\mathbf{v}$  is in fact the maximal data piling direction. Let  $\mathbf{w} \in \mathbb{R}^d$  be another vector with unit norm that gives complete data piling (i.e.  $\mathbf{w} \in P$ ). In particular, there exists  $a_v, a_w, b_v, b_w \in \mathbb{R}$  with  $a_v, a_w > 0$  such that

$$\mathbf{v}^T \mathbf{x}_i = a_v y_i + b_v \text{ for each } i = 1, \dots, n.$$

$$\mathbf{w}^T \mathbf{x}_i = a_w y_i + b_w \text{ for each } i = 1, \dots, n.$$

Assume for the sake of contradiction that  $\mathbf{w}$  projects the data possibly further apart than  $\mathbf{v}$  does. In particular assume that  $a_w \geq a_v$ .

Since  $\{\mathbf{t}_i\}_1^d$  is a basis we can write

$$\mathbf{w} = \sum_{i=1}^d \alpha_i \mathbf{t}_i.$$

Next compute the dot products with the data. For any  $j = 1, \dots, n$ ,

$$\mathbf{w}^T \mathbf{x}_j = \left( \sum_{i=1}^{n-1} \alpha_i \mathbf{t}_i \right)^T \mathbf{x}_j + \alpha_n \mathbf{t}_n^T \mathbf{x}_j + \sum_{i=n+1}^d \alpha_i \mathbf{t}_i^T \mathbf{x}_j.$$

Recall the basis vectors  $\mathbf{t}_{n+1}, \dots, \mathbf{t}_d$  are orthogonal to the data points so the third term in the sum is zero. Furthermore, the dot product of  $\mathbf{t}_n$  with each data point is a constant. Thus we now have

$$\mathbf{w}^T \mathbf{x}_j = \left( \sum_{i=1}^{n-1} \alpha_i \mathbf{t}_i \right)^T \mathbf{x}_j + \alpha_n c, \text{ for all } j = 1, \dots, n.$$

Thus we can see the vector

$$\mathbf{w}' = \sum_{i=1}^{n-1} \alpha_i \mathbf{t}_i$$

also gives complete data piling. However this vector lies in  $A$  since it is a linear combination of the first  $n-1$  basis vectors. We have shown that there is only one direction in  $A$  with complete data piling thus  $\sum_{i=1}^{n-1} \alpha_i \mathbf{t}_i \propto \mathbf{v}$ . In particular, for some  $\alpha > 0$

$$\sum_{i=1}^{n-1} \alpha_i \mathbf{t}_i = \alpha \mathbf{v}.$$

So we now have

$$\mathbf{w}' = \alpha \mathbf{v} + \alpha_n \mathbf{t}_n.$$

Recall  $\|\mathbf{v}\| = \|\mathbf{w}\| = 1$  and  $\mathbf{t}_n$  is orthogonal to  $\mathbf{v}$  by construction. Therefore  $\alpha^2 + \alpha_n^2 = 1$ . In particular if  $\alpha_n > 0$  then  $\alpha < 1$ .

Let  $\mathbf{x}_+$  and  $\mathbf{x}_-$  be any point from the positive and negative class respectively. By construction we have

$$\mathbf{v}^T (\mathbf{x}_+ - \mathbf{x}_-) = a_v.$$

$$\mathbf{w}^T (\mathbf{x}_+ - \mathbf{x}_-) = a_w.$$

However expanding this last line we get

$$\mathbf{w}^T (\mathbf{x}_+ - \mathbf{x}_-) = (\alpha \mathbf{v} + \alpha_n \mathbf{t}_n)^T (\mathbf{x}_+ - \mathbf{x}_-)$$

$$\mathbf{w}^T (\mathbf{x}_+ - \mathbf{x}_-) = \alpha \mathbf{v}^T (\mathbf{x}_+ - \mathbf{x}_-) + \alpha_n \mathbf{t}_n^T (\mathbf{x}_+ - \mathbf{x}_-).$$

But  $\mathbf{t}_n^T \mathbf{x}_+ = \mathbf{t}_n^T \mathbf{x}_- = c$  so the last term is zero. Thus we now have

$$\mathbf{w}^T (\mathbf{x}_+ - \mathbf{x}_-) = \alpha a_v.$$

Thus

$$\alpha a_v = a_w.$$

However unless  $\mathbf{w} = \mathbf{v}$  (so  $\alpha_n = 0$ ) we have  $0 < \alpha < 1$ . Therefore  $a_w < a_v$  contradicting the assumption that  $a_w \geq a_v$ . Therefore  $\mathbf{v}$  is the maximal data piling direction.  $\blacksquare$

## References

- Jeongyoun Ahn and J. S. Marron. The maximal data piling direction for discrimination. *Biometrika*, 97(1):254–259, 2010. ISSN 0006-3444. doi: 10.1093/biomet/asp084. URL <http://dx.doi.org/10.1093/biomet/asp084>.
- Jeongyoun Ahn, Myung Hee Lee, and Young Joo Yoon. Clustering high dimension, low sample size data using the maximal data piling distance. *Statistica Sinica*, pages 443–464, 2012.
- Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.
- Trevor Hastie, Saharon Rosset, Robert Tibshirani, and Ji Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5(Oct): 1391–1415, 2004.
- Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. A practical guide to support vector classification. 2003.
- Dougssoo Kaown. *A New Algorithm for Finding the Minimum Distance Between Two Convex Hulls*. dissertation, University of North Texas, 2009.
- Myung Hee Lee, Jeongyoun Ahn, and Yongho Jeon. Hdlss discrimination with adaptive data piling. *Journal of Computational and Graphical Statistics*, 22(2):433–451, 2013.
- James Stephen Marron, Michael J Todd, and Jeongyoun Ahn. Distance-weighted discrimination. *Journal of the American Statistical Association*, 102(480):1267–1271, 2007.
- Di Miao. *Class-Sensitive Principal Components Analysis*. PhD thesis, 2015.
- Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.
- Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- Tung Pham. *Some Problems in High Dimensional Data Analysis*. dissertation, University of Melbourne, 2010.
- Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.
- John Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.
- Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.